



NATIONAL
CENTER *for* ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington



Estimating Test-Score Growth with a Gap Year in the Data

Ishtiaque Fazlul
Cory Koedel
Eric Parsons
Cheng Qian

Estimating Test-Score Growth with a Gap Year in the Data

Ishtiaque Fazlul
University of Missouri

Cory Koedel
CALDER
University of Missouri

Eric Parsons
University of Missouri

Cheng Qian
University of Missouri

Contents

Contents	i
Acknowledgments.....	ii
Abstract	iii
1. Introduction.....	1
2. Methods	3
2.1 Growth Models	3
2.2 Gap-Year Simulation	7
3. Data.....	8
4. Results.....	9
4.1 Assessing the Alignment Between Gap-Year and Full-Data Growth Estimates	9
4.2 Factors that Predict Changes in Growth Rankings Induced by the Gap Year	13
5. Extension (2-Year Gap)	16
6. Discussion & Conclusion.....	18
References.....	20
Tables.....	21
Appendix A.....	26
Appendix B	B1

Acknowledgments

We thank the Missouri Department of Elementary and Secondary Education for data access and gratefully acknowledge financial support from the Fordham Institute and CALDER, which is funded by a consortium of foundations (for more information about CALDER funders, see www.caldercenter.org/about-calder). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of our funders, the Missouri Department of Elementary and Secondary Education, or the institutions to which the author(s) are affiliated. All errors are our own.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Estimating Test-Score Growth with a Gap Year in the Data

Ishtiaque Fazlul, Cory Koedel, Eric Parsons, Cheng Qian
CALDER Working Paper No. 248-0121
January 2021

Abstract

We evaluate the feasibility of estimating test-score growth with a gap year in testing data, informing the scenario when state testing resumes after the 2020 COVID-19-induced test stoppage. Our research design is to simulate a gap year in testing using pre-COVID-19 data—when a true test gap did not occur—which facilitates comparisons of district- and school-level growth measures that are estimated with and without a gap year. We find that growth estimates based on the full data and gap-year data are generally similar, and our results highlight an advantage of using comprehensive growth models with rich controls for student and schooling circumstances. With the caveat that there is looming uncertainty about which students will be tested in public schools when testing resumes, and how they will be tested, our findings establish the potential for estimating useful growth measures with a gap year in testing.

1. Introduction

Test-score growth is a commonly-used evaluation tool in education research and policy applications.¹ The abrupt cancellation of state testing in spring-2020 due to COVID-19 has generated a gap year in test data, a consequence of which is that it will not be possible to estimate traditional growth measures in 2021. Given this data condition, we assess the potential for reliably estimating test-score growth over the two-year period from spring-2019 to spring-2021.

Our methodological approach is to simulate a gap year in testing in a year preceding COVID-19. Specifically, we build a data panel spanning the school years 2016-17, 2017-18, and 2018-19, and censor the data as if the 2017-18 test was never administered. We estimate value-added models using the artificially-censored data and compare the output to analogous output obtained using the full, uncensored data panel over the same two-year period. These comparisons allow us to assess the accuracy of gap-year growth estimates relative to the full-data condition.

We focus primarily on determining the accuracy with which we can estimate test-score growth for districts and schools. District- and school-level growth estimates are inputs in many state accountability plans under ESSA and can be used to assess gaps in learning during the pandemic more broadly. Moreover, the presence of accurate district- and school-level growth estimates implies that credible, growth-based analyses of district- and school-level interventions during the pandemic will be possible (subject to other standard evaluation concerns). A companion policy report documents our high-level findings (Fazlul, Koedel, Parsons, and Qian, 2021)—this article expands on these findings and provides technical details for researchers interested in estimating test-score growth when testing resumes.

¹ As a part of their accountability plans under the Every Student Succeeds Act (ESSA), forty-seven states plus Washington DC indicate using some form of growth measure for elementary and middle schools. For high schools, 20 states indicate using student growth measures for accountability (source: Education Commission of the States, retrieved on 12.21.2020 at <http://ecs.force.com/mbdata/mbQuest5E?rep=SA172>). The teacher-evaluation landscape is rapidly evolving and it is harder to get current national numbers, but the reliance on growth measures to evaluate teachers is generally waning (e.g., see Dee, James, and Wyckoff, 2019; National Council on Teacher Quality, 2019). Researchers also regularly rely on growth models to assess the efficacy of education interventions.

We find that estimates of test-score growth using the gap-year data are highly correlated with estimates that use all of the data. Specifically, the correlations are consistently around 0.90 for districts, and range between 0.84-0.88 for schools, across five different growth-model specifications in two subjects (Math and English Language Arts). We also extend our analysis to briefly consider a scenario where testing is further postponed to 2022. We do not believe it will be feasible to estimate test-score growth for individual schools with a 2-year test gap, but reasonably informative district-level growth estimates could still be estimated if this happens.

Though we find broad similarity between the results obtained under the gap-year and full-data conditions, the estimates are not identical. For the differences that exist, we investigate their sources and identify two primary factors. First, the cohorts of students used to estimate growth in the two scenarios only partially overlap. If we force the gap-year and full-data growth models to use the same student cohorts, the correlations reported in the previous paragraph rise by about 0.05. Second, the remaining discrepancies between the gap-year and full-data estimates are the result of what we refer to as data and modeling variance—i.e., they arise because the gap-year model estimates growth from period $t-2$ to t , whereas the full-data analog sums single-year estimates of growth from $t-2$ to $t-1$, and $t-1$ to t . This generates small differences in the predictors of contemporary achievement and their coefficients. We rule out other sources of the discrepant results due to the gap year. Most significantly, conditional on cohort-alignment, sampling variance at the individual student level does not meaningfully impact estimated growth for districts and schools using the gap-year data.

We also examine the extent to which differences in growth rankings caused by a gap year can be systematically predicted by observable district and school characteristics. We find that most of the variance in growth-ranking changes is not explained by observable characteristics. However, in sparser growth models they explain a non-negligible share of the variance—up to 25 percent. In our richest growth specification, the explained variance falls dramatically into the range of 1-5 percent, highlighting a desirable property of that specification.

Our findings provide a foundation for efforts to restart growth modeling beginning in spring-2021 by establishing the estimation consequences of a gap year in test data. The fact that we find relatively small impacts of the gap year using pre-pandemic data—absent other major disruptions—is useful because it suggests that growth estimates from the pandemic period are not likely to be (meaningfully) confounded by statistical issues stemming from the missing year of data. However, an important caveat to our analysis is with regard to possible changes in the composition of test-takers in public schools when testing resumes. For all of our analysis we assume that after the test gap, comprehensive testing will resume as before. This is an instructive starting point for thinking about the prospects for growth modeling with a gap year, but as of the writing of this article there is great uncertainty about test coverage in spring-2021 in the context of the ongoing pandemic. A related source of uncertainty is the potential for a change to the composition of testing modes (e.g., in-person versus online) and whether tests across different modes can be modeled together effectively. Because the uncertainty along these dimensions is so great and conditions vary so much across locales (e.g., see Donaldson and Diemer, 2021; Goldhaber et al., 2020), we do not attempt to address this issue directly in our simulations. Noting this limitation, we view our work on the technical implications of the gap year as a crucial first step toward re-starting our growth modeling infrastructure post-pandemic. Our findings provide a foundation for real-time research to address testing composition issues as more information is revealed about who is tested, and how, in states and districts across the U.S. when testing resumes.²

2. Methods

2.1 Growth Models

We estimate five different value-added models (VAMs) in math and ELA, described in broad terms in Table 1. For each model-subject combination, we estimate the model separately to

² Several states have recently indicated that testing will occur in spring 2021, but without accountability (e.g., Missouri and Texas). The lessons from our analysis will apply to growth-based analysis of the 2021 test data regardless of whether accountability is enforced.

recover estimates of district and school value-added. The models differ in terms of structure and the variables included. Regarding the former, we estimate both “one-step” and “two-step” VAMs.

Our full specification for the one-step VAM used to estimate school value-added is shown by equation (1):

$$Y_{ijkmt} = \alpha_0 + \mathbf{Y}_{imt-1}\boldsymbol{\alpha}_1 + \mathbf{X}_{it}\boldsymbol{\alpha}_2 + \gamma_j + \varphi_s + e_{ijkmt} \quad (1)$$

Equations (2) and (3) show the full specification for the two-step VAM:

$$Y_{ijkmt} = \beta_0 + \mathbf{Y}_{imt-1}\boldsymbol{\beta}_1 + \mathbf{X}_{it}\boldsymbol{\beta}_2 + \bar{\mathbf{Y}}_{mst-1}\boldsymbol{\beta}_3 + \bar{\mathbf{Y}}_{mkt-1}\boldsymbol{\beta}_4 + \bar{\mathbf{X}}_{st}\boldsymbol{\beta}_5 + \bar{\mathbf{X}}_{kt}\boldsymbol{\beta}_6 + \psi_j + \varepsilon_{ijkmt} \quad (2)$$

$$\varepsilon_{ijkmt} = \pi_s + u_{ijkmt} \quad (3)$$

In equations (1) and (2), Y_{ijkmt} is standardized test score of student i in grade j , attending school s , in district k , for subject m , and year t . \mathbf{Y}_{imt-1} is a vector of lagged-test-score controls containing four elements. The first element is the same-subject lagged test score, which is required of all students for inclusion in each subject-specific model (i.e., math or ELA). The second element is the lagged off-subject score—in the math models we include the lagged ELA score, and in the ELA models we include the lagged math score. We include students who are missing the lagged off-subject test score (but have the lagged same-subject score) in the models by imputing the missing score to the mean. Correspondingly, we also include an indicator variable that takes a value one if the score is missing and zero otherwise. Finally, we add an interaction between the missing indicator for the off-subject lagged test score and the lagged same-subject score, which improves estimation efficiency by allowing the model to rely more heavily on same-subject lagged performance to predict current performance for students who are missing the off-subject lagged score. In equation (2), the vectors $\bar{\mathbf{Y}}_{mst-1}$ and $\bar{\mathbf{Y}}_{mkt-1}$ include school and district average values of the lagged test-score variables. In the uncensored data condition without the gap year, the lagged-score controls are included as described in this paragraph. In the gap-year models, $t-1$ scores are unavailable and $t-2$ scores are substituted.

The vector \mathbf{X}_{it} contains student characteristics. We include indicator variables for race/ethnicity, gender, free and reduced-price lunch (FRL) status, English language learner (ELL)

status, whether the student has an individualized education program (IEP), and mobility status (i.e., an indicator for whether the student changed schools mid-year). We also include the school and district shares of these variables in the vectors \bar{X}_{st} and \bar{X}_{kt} in equation (2). γ_j and ψ_j are grade fixed effects. As written, in equations (1) and (3), φ_s and π_s are school fixed effects. These are our estimates of school value added. Note that when we estimate district value-added, we replace the school fixed effects with district fixed effects (i.e., with subscripts “ k ” instead of “ s ”), re-estimate the models, and recover these parameter estimates instead. e_{ijkmt} , ε_{ijkmt} , and u_{ijkmt} are the error terms.

The equations shown in (1)-(3) describe the full versions of the one-step and two-step models and are labeled as “Model 3” and “Model 5” in Table 1. Models (1) and (2) in Table 1 are sparse versions of the one-step and two-step models—they include only the Y_{imt-1} vector and the grade fixed effects. Model (4) is a two-step model that includes all of the information in the full one-step model shown in equation (1)—i.e., it includes all student-level controls but excludes all district- and school-aggregated information. Note that the school- and district-aggregate coefficients are not separately identified in a one-step model because there is no within-unit (school or district) variation in the aggregate covariates. This is why we do not estimate a one-step model with these controls. The two-step model “resolves” the identification problem by estimating the parameters sequentially. It is beyond the scope of the current paper to go into details on the technical and policy tradeoffs of the various models, but Ehlert et al. (2016) and Parsons et al. (2019) provide conceptual and technical arguments for why a 2-step model with rich controls along the lines of Model (5) is desirable.³

We attribute student growth to the contemporary school or district in all models as a baseline condition. This is the common approach under normal circumstances—i.e., growth from

³ Interested readers can find discussions of the technical and policy tradeoffs of the various models in the following articles, among others: Ehlert, Koedel, Parsons, and Podgursky (2016); Goldhaber, Walch, and Gabele (2014); Guarino, Reckase, and Wooldridge (2015); Guarino, Reckase, Stacy, and Wooldridge (2015); Kane, McCaffrey, Miller, and Staiger (2013); Koedel, Mihaly, and Rockoff (2015); and Parsons, Koedel, and Tan (2019). Most of these papers focus on estimating teacher value-added, although the general insights apply to other levels of value-added as well.

year- $(t-1)$ to year- t is attributed to the year- t school or district. In the gap year model, this is a potential concern because there is extra mobility during the gap year. We examine the sensitivity of gap-year model output to adjustments for student mobility over the course of our analysis.

The last estimation issue that merits brief mention is shrinkage. All of our estimates are shrunken toward the mean using the following procedure described in Koedel, Mihaly, and Rockoff (2015), which is implemented in two steps. First, for each school or district estimate we produce an estimate-specific shrinkage factor, α . For each school s , the shrinkage factor is written as:

$$\alpha_s = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\lambda}_s} \quad (4)$$

In the formula, $\hat{\sigma}^2$ is an estimate of the variance of true value-added across schools in the sample (after netting out estimation error), and $\hat{\lambda}_s$ is the estimation-error variance of the estimate for school s .⁴ These shrinkage factors can be thought of as individual school (or district) reliability ratios that reflect the precision of each estimate in the context of the total true variance in value added.

With the shrinkage factors in hand, the final, shrunken value-added estimates are calculated as (again, the formula for schools is shown but the formula for districts is analogous):

$$\tilde{\varphi}_s = \alpha_s \hat{\varphi}_s + (1 - \alpha_s) \bar{\varphi} \quad (5)$$

where $\hat{\varphi}_s$ is estimated value-added for school s from the regression and $\bar{\varphi}$ is average value-added across all schools. Equation (5) embodies the intuitive idea that as the estimate for any individual school s becomes less precise, as measured by α_s , we put more weight on the prior that the school is average in terms of value-added.⁵

⁴ We estimate $\hat{\lambda}_s$ as the square of the standard error of the value-added coefficient for school s . Note that for the estimates from the 2-step models, we use the standard errors from the second step in these calculations. This is a simplification because it ignores estimation error in the first step. In omitted results, we confirm that the practical implications of this simplification are ignorable by comparing this approach to a comprehensive approach in which we bootstrap the entire two-step procedure to account for estimation error in both the first and second steps.

⁵ There are two somewhat common growth-model approaches not directly covered by our analysis: “student growth percentiles” and EVAAS®. With regard to the former, although we do not estimate student growth percentiles directly, Ehlert, Koedel, Parsons, and Podgursky (2016) show that a linear model using similar information produces similar results, which implies that our results using the sparse VAMs should be a reasonable approximation of results that would be obtained using student growth percentiles. We also do not evaluate EVAAS®, which is a semi-

2.2 Gap-Year Simulation

Using each model described above, we estimate value-added with and without censoring the data to simulate a gap year in testing. We begin by using the uncensored data to estimate two consecutive value-added estimates for each unit (either a school or a district) with data from 2016-17 to 2017-18, and 2017-18 to 2018-19. We then sum the two single-year estimates to produce an estimate of value-added over the 2-year period to replicate how a typical system would estimate value-added over two years, assuming no data were missing.

Next, to simulate the gap year in testing, we censor the 2017-18 test data and directly estimate value-added over the 2-year period, using data from 2016-17 and 2018-19. This scenario is meant to reflect the data condition when (and if) testing resumes in spring 2021; i.e., the condition of a gap year between the 2018-19 and 2020-2021 tests. By comparing the “full data” scenario to the “gap year” scenario, we can assess the extent to which the gap-year models recover accurate estimates of test-score growth over the two-year period.⁶

We also briefly extend our analysis to simulate the presence of two consecutive gap years in testing—this would be the scenario if testing does not resume until spring-2022. For this extension, we bring in an additional year of earlier data from 2015-16, censor the test data in our panel in 2016-17 *and* 2017-18, and calculate growth from 2015-16 to 2018-19. We then compare growth estimated over the three-year period to the analogous “full data” condition, where three-year growth is calculated as the sum of annual growth estimates from 2015-16 to 2016-17, 2016-17 to 2017-18, and 2017-18 to 2018-19.

proprietary growth model administered by the SAS Institute, but note that Vosters, Guarino, and Wooldridge (2018) find a high level of agreement between SAS’s univariate response model and other value-added approaches.

⁶ Our research design builds on a large existing literature showing that estimates of test score growth from value-added models using the typical full-data condition provide reliable information (e.g., see Chetty, Friedman, and Rockoff, 2014; Deming, 2014; Kane et al., 2013; Koedel, Mihaly, and Rockoff, 2015).

3. Data

We use administrative microdata from Missouri covering all students tested in grades 3-8 in math and ELA during the school years 2015-16 to 2018-19. Hereafter, we identify school years by the spring year—e.g., 2018-19 as 2019.

We standardize student test scores throughout by grade-subject-year and estimate growth for all districts and schools with at least 10 tested students. When we correlate and otherwise compare growth estimates using the full data and gap-year data, the comparisons are restricted to districts and schools that meet the size threshold in both data conditions. Only very small Missouri districts and schools are omitted from our analysis due to the sample-size restriction.⁷

We do not expect contextual features of Missouri to limit the generalizability of our findings in most respects. That said, two aspects of the Missouri data merit brief attention. First, Missouri changed its math and ELA tests once each between 2016 and 2019. Backes et al. (2018) study the impact of test-regime changes on value-added estimates in math and ELA across multiple states and find that such changes typically do not affect model performance substantively. Moreover, we have performed internal diagnostic work using the Missouri data specifically that supports this inference.⁸

Second, Missouri has a high ratio of districts to students. Said another way, Missouri is a “small district” state. Growth estimates for smaller districts will be more sensitive to data changes because they have fewer students to balance out the sampling variance that the data changes create. These data changes can be of two types. First is the imperfect overlap of the samples between the full-data and gap-year scenarios, both at the cohort and individual-student levels.⁹ Second, even with perfect overlap of students in the full-data and gap-year scenarios,

⁷ About 1-2 percent of Missouri districts and schools are excluded for this reason (and fewer than 0.10 percent of students, noting that the districts and schools that are excluded are small).

⁸ For example, at the student level, the predictive value of prior achievement as the testing regime changes is stable.

⁹ As an example of imperfect cohort overlap, note that students in 7th grade in 2017 and 9th-grade in 2019 will be part of the analysis in the full data scenario but not in the gap year scenario. This is because in 2019 when testing data are again available, the student will be outside of the tested gradespan. Imperfect student overlap within cohorts can also occur—e.g., a 4th grader in 2017 could miss her test in that year but take the tests in the 5th and 6th grades in 2018 and 2019, in which case she would be partly included in the full data scenario but not the gap year scenario.

differences in the same students' test scores in the $t-1$ and $t-2$ years can affect the growth estimates. We conduct a subsample analysis for the 100 largest districts in Missouri—in which these data changes are less impactful owing to their size—to produce results that are more likely to generalize to states with larger school districts.

Table 2 summarizes our data in terms of students, schools, and districts.

4. Results

4.1 Assessing the alignment between gap-year and full-data growth estimates

Under a variety of conditions we estimate district- and school-level growth using the full data, and then using the censored data as if the 2018 test was not administered, and compare the results by estimating the correlation between the growth estimates. Each cell in Table 3 shows one such correlation between school- or district-level growth estimates, with and without the data censoring, defined by three dimensions: (1) the subject (math or ELA) and model (Models 1-5) indicated by the column, (2) the level of the analysis (district or school) indicated by the two horizontal panels, and (3) the precise data and evaluation condition, identified by the rows within each horizontal panel.

Our baseline findings for districts and schools are reported in the first row of each horizontal panel. The two key features of the baseline condition, both of which we relax subsequently, are (a) we compare the gap-year and full-data results using all available data in each condition, and (b) we assign growth over the previous period—be it one ($t-1$) or two ($t-2$) years—to the year- t district or school, which is the business-as-usual approach in the absence of a gap year.

The baseline results for districts show that the gap-year estimates are highly correlated with the full-data estimates in both subjects. The correlations are consistently around 0.90 and slightly higher in math. The correlations are a little lower for individual schools—in the range of

0.84-0.88 across models and subjects—but substantively similar.¹⁰

A high-level takeaway from the baseline correlations is that they indicate a strong correspondence between growth estimated with and without the gap year, regardless of level of analysis, growth model, or subject. In Appendix Table A1, we provide complementary transition matrices corresponding to the baseline correlations. Reflecting the fact that research and policy interest is often concentrated in the tails of the distribution, the transition matrices examine the persistence of district and school placements in the “bottom 10 percent,” “middle 80 percent,” and “top 10 percent” of growth rankings with and without the gap year. Mirroring the high correlations in Table 3, the transition matrices consistently show that most districts and schools (about 85-88 percent) remain in the same ranking category regardless of whether the full data or gap-year data are used. Moreover, as expected, the districts and schools that change categories are relatively close to the 90th- and 10th-percentile cutoffs, on average; among these districts and schools, the average value of the percentile ranking change caused by the gap year of data is about 10 percentile points—e.g., a move from the 85th to 95th percentile.

Noting the baseline correlations are generally high, one might still wonder why they are not even higher. After all, both the gap-year and full-data models aim to recover growth estimates over the same two-year period. Understanding what factors drive differences between the estimates is important for understanding the limitations of using gap-year data.

In the second and third rows in each panel of Table 3, we explore the extent to which changes in the analytic sample between the gap-year and full-data models can explain differences in the results. First, in the second row of each panel, we force the gap-year and full-data models to be estimated on the same cohorts of students. In the baseline condition, the full-data models include some cohorts who are not represented in the gap-year models. As an

¹⁰ The SAS Institute (undated) reports correlations for growth estimates for districts and schools from its proprietary TVAAS® Multivariate Response Model (MRM) with and without a gap year. SAS reports much higher correlations at both levels (0.99), purportedly using a similar research design. We were surprised by this result, and in subsequent correspondence with SAS researchers we learned that the correlations they report are not analogous to the correlations we report here. Our interpretation is that the analysis reported on by the SAS Institute (undated) is not directly informative about the accuracy of gap-year growth estimates relative to the full-data counterfactual.

example, consider a student in the third grade during the gap year, which for us is 2018. Her growth contributes to estimates from 2018 to 2019 in the full data condition, but because she is outside of the tested range prior to 2018 (i.e., in 2017 she is in the second grade), her growth cannot be assessed with the gap-year model. A similar problem arises for students in the eighth grade during the gap year, who age out of the testing window before testing resumes.

When we force cohort alignment between the models, the correlations in all scenarios rise markedly, on the order of about 0.05 off of the already high baseline values. This indicates that cohort misalignment between the gap-year and full-data conditions accounts for a substantial fraction of the result discrepancies. This finding is not likely to be useful for policy because in the presence of a true gap year, the missing cohorts will simply not have data. However, it is instructive about why the growth estimates differ, which can be valuable for researchers and policymakers as they consider various uses of gap-year growth data after COVID-19.

Next, in the third row in each panel of Table 3 we further align the samples across the gap-year and full-data conditions by using the exact same students to estimate the models. That is, conditional on cohort alignment, we further exclude all students within the matched cohorts that do not have a test score for all three years (2017, 2018, and 2019). The results show that conditional on matching cohorts, matching the exact student samples has a negligible effect on the comparability of gap-year and full-data growth estimates. The correlations do increase when we fix the samples, but the increase is very small, and in some cases is not detectable up to the hundredth decimal place in the correlations.

Another way that the gap-year and full-data models differ is in how they treat mobile students. To illustrate, consider a student who attends District A in 2017 and 2018, but district B in 2019. In the business-as-usual VAM, her growth from 2017 to 2018 will be attributed to District A, and her growth from 2018 to 2019 will be attributed to District B. However, using the convention of assigning growth to the contemporary district in the gap-year model from 2017 to 2019, her growth over the full 2-year period will be attributed to District B.

We assess the extent to which two different mobility-based adjustments to the gap-year model improve its performance. First, we drop students from the gap-year model who were not enrolled in the same district (or school) in period $t-1$ and t —i.e., in 2018 and 2019 in our censored dataset. These students only attended the contemporary district (or school) for one of the two years over which gap-year growth is estimated, meaning that their full-period growth is partly misattributed using the convention of assigning growth to the contemporary location. In the second mobility modification, we retain all mobile students in the gap-year dataset but assign 50 percent weight to the districts (or schools) attended in 2018 and 2019, respectively.¹¹

The results in rows 4 and 5 of each panel of Table 3 show correlations between the gap-year and full-data models, after making the mobility adjustments to the gap-year models. The correlations otherwise maintain the baseline evaluation conditions, so the effects of the mobility adjustments can be inferred by comparing the results to the results in row 1. For districts, neither mobility adjustment results in an improvement in the performance of the gap-year model. In fact, the adjustment where we drop mobile students altogether (weakly) reduces the ability of the gap-year model to recover the full-data growth estimates. The reason is that the lost data reduces efficiency, offsetting any (very modest) gains owing to the reduced misattribution of mobile students' growth.

For schools, the strategy of dropping the data for movers also performs (weakly) worse for the same reason. However, the 50-50 weighting strategy modestly improves estimation accuracy in the gap-year model. A reason the results differ between districts and schools—albeit only slightly—is that there are many more school than district movers during the gap year.¹²

In Appendix Table A2 we replicate the analyses in Table 3 for the subsample of the 100 largest districts in MO, noting that these findings will be more generalizable to “large district”

¹¹ We use the 2018 test data to assign students to districts and schools in 2018. With a true gap year, test data would be unavailable, but this could be achieved using enrollment records instead.

¹² A factor that drives higher school mobility, in addition to the fact that school catchment areas are smaller than district catchment areas, is that there are many more “structural” school movers. A structural school move is a move that occurs because a school's gradespan has ended, e.g. due to a transition from elementary to middle school.

states. The findings are substantively similar to our results for districts in Table 3 in all facets, although the baseline correlations are higher owing to the larger sample sizes.

In summary, Table 3 shows that cohort-misalignment is the single largest, tractable factor that drives down the baseline correlations between the gap-year and full-data growth estimates. With cohort alignment, the correlations are consistently around 0.95 and 0.90 across models in the district- and school-level analyses, respectively, and even higher in the large-district subsample (Appendix Table A2). In the school-level models, differences in how the full-data and gap-year models attribute growth for mobile students is also a small contributing factor. We are left to conclude that the remaining discrepancies arise from data and modeling variance.¹³ Again, this variance stems from the fact that we model growth from $t-2$ to $t-1$ and from $t-1$ to t in the full-data models, and directly from $t-2$ to t in the gap-year models. Individual students' $t-1$ and $t-2$ test scores are different (data variability), and the model coefficients on the $t-1$ and $t-2$ coefficients are different, which in turn can affect other coefficients in the models as well (modeling variability). As unit-level (district or school) sample sizes become large, the effect of the data variability shrinks, but the effect of modeling variability does not.

4.2 Factors that predict changes in growth rankings induced by the gap year

Next, we assess whether observable district and school characteristics predict ranking changes between the gap-year and full-data models under the baseline estimation conditions. Tables 4 and 5 show results from regressions where the dependent variable is the difference in growth rankings between the gap-year and full-data models—i.e., we estimate each model separately, assign districts and schools percentile ranks based on the growth estimates, and subtract the full-data percentile from the gap-year percentile. The independent variables are district and school characteristics including the 2017 same-subject achievement level, the number of test takers, and student shares by race-ethnicity, gender, FRL, English as a second

¹³ We also shrink each estimate separately in the full-data model, and this has the potential to generate small differences between conditions because the gap-year output is only shrunk once. However, in results omitted for brevity we verify our findings are nearly identical without shrinkage, ruling out this procedural difference as a driver of divergent results between models.

language (ESL), participation in an individualized education program (IEP), and student mobility (in particular, the share of tested students who experienced a mid-year school move). All of the independent variables are standardized to have a mean of zero and a variance of one—within the district or school distribution depending on the level of analysis—which allows the coefficients to be interpreted in (common) standard-deviation units throughout.

We begin by focusing on the R-squared values in Tables 4 and 5, which give a summary indication of the predictive power of observable characteristics over gap-induced changes to growth rankings. For the growth estimates from Models 1-4, the R-squared values indicate a non-negligible fraction of the variance in ranking changes can be explained by observable district and school characteristics— about 14-25 percent for districts and 10-18 percent for schools. Alternatively, in Model 5, our fullest specification, observable district and school characteristics explain almost none of the variance in ranking changes—about 4-5 percent for districts and 1-4 percent for schools.

The primary predictor of the rank changes in all models and subjects is the 2017 achievement level. The consistently negative coefficients on that variable using the estimates from Models 1-4 indicate that higher-achieving districts and schools are adversely affected in growth rankings by the presence of the gap year, compared to the full-data analog. The magnitudes of the relationships are moderate, with a one-standard-deviation increase in the 2017 achievement level corresponding to a ranking reduction of about 5-8 percentile points. Noting that the 2017 achievement level is a broad indicator of socioeconomic advantage, it also bears mentioning that the coefficients on some of the other control variables in the multivariate regressions temper the relationship between advantage and lower rankings, on net. As an example, take Model 1 in Table 4 for math. Negative ranking changes in that model due to the gap year are also associated with higher percentages of underrepresented minority students, FRL-eligible students, IEP students, and geographically mobile students. Still, on the whole, the lagged test-score coefficient dominates all of these, and the end result is that moving from the

full data condition to the gap-year data condition in Models 1-4 systematically lowers estimated growth for socioeconomically advantaged districts and schools.¹⁴

A theoretical explanation for the findings from Models 1-4 is provided in Appendix B. In short, the appendix shows the findings are consistent with the presence of modest omitted variables bias in the underspecified VAMs. This bias is fully compounded in the consecutive single-year estimates used in the full-data scenario but partially attenuated in the gap-year estimates. The bias explanation is consistent (conceptually and directionally) with the bias documented in underspecified VAMs in Parsons, Koedel, and Tan (2019) and implies that the gap-year estimates are less biased than their full-data counterparts. We caution that this does not mean that the gap-year estimates from the underspecified VAMs are preferred because they have other limitations, most notably in terms of coverage and sample sizes. However, they are less-biased.¹⁵

The finding that changes to the growth rankings based on Model 5 are not meaningfully explained by observable district and school characteristics, combined with the derivations in Appendix B, is consistent with that model producing the least-biased growth estimates. However, the evidence is not conclusive because Model 5 may be “overcorrecting” for student and school circumstances. Previous research suggests that overcorrection bias in fully-controlled 2-step VAMs, like Model 5, is more problematic in theory than in reality, but it is beyond the scope of the present article to delve into these details further. We refer interested readers to

¹⁴ To assess the net effect more directly, we also estimate versions of the models shown in Tables 4 and 5 that only include a single covariate: the lagged aggregate test score. The influence of student demographics correlated with test scores that work in the opposite direction are absorbed by the coefficient on the lagged aggregate test score in these models, and as a result its magnitude is about 20 percent smaller in these supplementary regressions than what we show in Tables 4 and 5. These results are omitted for brevity but available upon request.

¹⁵ In the interest of scientific transparency, we did not anticipate the finding that observable characteristics would systematically predict ranking changes caused by the gap year in any of the models *ex ante*, and the theoretical explanation provided in Appendix B was developed *ex post*.

Ehlert, Koedel, Parsons, and Podgursky (2016) and Parsons, Koedel, and Tan (2019) for more information.¹⁶

The results in Tables 4 and 5 highlight an advantage of using the rich, two-step specification exemplified by Model 5 for growth-based research, both in general and during the pandemic period. While Models 1-4 perform similarly to Model 5 in other respects (e.g., see Table 3), their use of weaker sets of control variables increases the potential for a confounding of growth with other district and school characteristics, which increases the likelihood of incorrect inference about which students have been affected by the pandemic and by how much.

5. Extension (2-year gap)

Thus far we have assumed that testing will resume in spring-2021. Here, we briefly consider the prospects for estimating growth for schools and districts if testing does not resume until spring-2022. If this were to happen, there would be a 2-year gap in testing between 2019 and 2022. In our data, we simulate this situation by adding a year to the front end of our data panel and further censoring the data to remove the 2017 test. That is, we expand the test gap period to include both 2017 and 2018 and estimate growth from 2016 to 2019, replicating the “2-year gap” condition that would arise if spring-2021 testing is cancelled.

In this scenario, our view is that school-level growth metrics cannot be feasibly estimated. This is because most schools would not have any students who take both the pre- and post-gap tests in the same building, which would require schools to cover four consecutive grades in the tested span (grades 3-8). For example, third-grade students in a K-5 school in the pre-gap year would be sixth graders in a new school after a 2-year gap.¹⁷ Complex and

¹⁶ Model 5 partials out variance in student growth explained by school and district characteristics in equation (2) before constructing the residual-based growth estimates in equation (3). Thus, it is built to produce growth estimates that are generally uncorrelated with measured district and school characteristics. Modest correlations are mechanically possible because the full-data models control for t-1 aggregate achievement whereas the gap-year models control for t-2 aggregate achievement, and indeed Tables 4 and 5 show that gap-induced ranking changes are not entirely uncorrelated with district and school characteristics. That said, the coefficients are much smaller in Model 5 than in Models 1-4, and their substantive implications are not meaningful.

¹⁷ The only somewhat common grade configuration in the 3-8 range that meets this criterion is K-6; K-5, 6-8, and 7-8 schools, among other configurations, fall short. Using the Common Core of Data from 2018-19, we estimate that just 27 percent of students enrolled in grades 3-8 in a U.S. public school attend a school with four consecutive grades in the 3-8 range.

assumptive models could theoretically recover estimates of test-score growth for individual schools even in the absence of “fully-contained” cohorts within schools to anchor the estimates, but without considerable validity testing, we do not view this as a promising strategy.¹⁸

Alternatively, district-level growth estimates with a 2-year gap can be feasibly estimated because most districts span four consecutive grades in the 3-8 span. Students transitioning across schools, as long as they stay in the same district, are not problematic for estimating test-score growth at the district level. Still, the extra year of missing data does present challenges, even for estimating district growth. The biggest challenge is that growth can be estimated for even fewer cohorts. Specifically, students in grades 3, 4, and 5 in the pre-gap year are the only students for whom an endpoint score would be available after a 2-year gap. Given that a lack of cohort overlap is a key driver of discrepancies in district growth estimates with and without a single gap year, a prediction is that with a 2-year test gap and fewer available cohorts, the discrepancies will be larger.

In Appendix Table A3 we partially replicate the analysis in Table 3 for districts using the 2-year gap scenario. Consistent with our expectation, with a 2-year test gap, the gap-year growth estimates from 2016-2019 are less correlated with estimates based on the full data (in this case, three years of summed, single-year estimates). The baseline correlations in Appendix Table A3 range from 0.78-0.84, compared to 0.88-0.91 in the case of a 1-year test gap in Table 3. The correlations are still large and positive, but they also indicate a larger degradation of information relative to the full-data case. Like in our analysis of the single-year gap, cohort alignment greatly improves agreement in the output between the gap-year and full-data conditions in Appendix Table A3, although the correlations are lower across the board with a 2-year gap.

¹⁸ This would involve assigning partial credit for student growth over the 2-year gap period depending on the share of time that each student spends in each school between tests. For example, growth estimates for separate K-5 and 6-8 schools could be constructed by leveraging the fact that third graders and fourth graders in the pre-gap year had different levels of exposure to the K-5 and 6-8 schools between tests. We also note that even if such an approach could be shown to be well-grounded scientifically (which is uncertain), the prospects for securing buy-in for such an approach for use in policy applications seem limited, especially if the growth metrics are to be used for accountability.

6. Discussion & Conclusion

Growth modeling is the most promising large-scale approach for assessing the impacts of the COVID-19 pandemic on student learning in K-12 schools. However, the pandemic has impacted our ability to measure test-score growth. This paper addresses the immediate and obvious concern that it will not be possible to estimate traditional year-to-year test score growth for individual students after testing resumes because scores from the previous year will be unavailable. Using pre-COVID-19 data and a research design based on a simulated gap year in testing, we evaluate the ability of growth estimates based on gap-year data to replicate growth estimates based on all of the data. We observe the latter because our analysis is based on a simulated, rather than real, gap-year scenario.

The fact that we conduct our analysis using pre-COVID-19 data is useful because it allows us to understand the technical consequences of estimating growth with a gap year in the absence of other disruptions. Across a range of models that are broadly representative of those used in research and policy applications, we show that gap-year growth estimates for districts and schools are highly correlated with estimates that would be obtained in a full-data condition if the gap year did not occur. For districts, correlations between gap-year and full-data growth estimates across models and subjects in Missouri are on the order of 0.90 (and as high as 0.95 for a subset of large districts), and analogous correlations for schools are in the range of 0.84-0.88. These findings suggest that gap-year growth estimates from the pandemic period are unlikely to be (meaningfully) confounded by statistical issues attributable to the gap year itself, and lend credence to their use as measures of student learning.

All of the growth models we consider perform similarly in the presence of the gap year along most dimensions. The one exception is in the extent to which growth-ranking changes caused by the gap year are systematically related to observable district and school characteristics. Our richest growth specification—a two-step value-added model with extensive controls—produces the most desirable output along this dimension. In Appendix B, we show this can be explained by the presence of greater omitted variables bias in the sparsely-specified VAMs,

which the gap-year models attenuate to some degree. Our fullest specification is appealing for assessing student learning trajectories during the pandemic because growth estimates are less likely to be confounded by the characteristics of districts and schools.

In a brief extension we also consider the potential for estimating test-score growth if there is a 2-year gap in testing. With a 2-year gap, we argue that it will be infeasible to produce growth metrics for most K-12 schools. District-level metrics could still be estimated. They will be less reliable compared to the full-data condition owing to the larger gap period, but will still contain useful information about test-score growth.

Finally, we conclude by noting that we view this work as just the first of two major steps in the larger process of rebuilding our capacity to measure test-score growth as we emerge from the COVID-19 pandemic. One way to describe our findings is that they document the *potential* for using gap-year data to estimate test-score growth. However, our analysis ignores the other problematic aspect of the pandemic—when testing resumes, it is not clear who will be tested and how (i.e., in-person or online). There is currently considerable variability in learning modes across districts within states, and across states, which will likely map to differential testing rates and testing modes for students this spring. For example, as of November 2020, 64 percent of districts in Michigan were operating fully in-person, versus just 13 percent in Washington state (Goldhaber et al., 2020). Test coverage, and selection into test coverage, is the other major analytic issue that will need to be addressed before we can begin to confidently apply growth models for use in understanding the learning impacts of the pandemic. At the time of writing this article, there are too many dimensions of uncertainty with respect to test coverage for us to feel confident in parameterizing useful partial-coverage simulations. Noting this outstanding question, our work provides a jumping off point for real-time research to resolve this issue as data from spring testing become available.

References

- Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62, 48-65.
- Chetty, R., Friedman, J.N., and Rockoff, J.E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9), 2593-2632.
- Dee, T. S., James, J., & Wyckoff, J. (2019). Is Effective Teacher Evaluation Sustainable? Evidence from DCPS. CEPA Working Paper No. 19-09. *Stanford Center for Education Policy Analysis*.
- Deming, D. J. (2014). Using school choice lotteries to test measures of school effectiveness. *American Economic Review*, 104(5), 406-11.
- Donaldson, K., & Diemer, A. (2021). Missouri Covid-19 reopening profile. St. Louis, MO: PRiME Center at St. Louis University.
- Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy* 30(3), 465-500.
- Fazlul, I. Koedel, C., Parsons, E., Qian, C. (2021). Bridging the COVID Divide: How States Can Measure Student Achievement Growth in the Absence of 2020 Test Scores. Washington, DC: Fordham Institute.
- Goldhaber, D., Imberman, S.A., Strunk, K., Hopkins, B., Brown, N., Harbatkin, E., and Kilbride, T. (2020). To What Extent Does In-Person Schooling Contribute to the Spread of COVID-19? Evidence from Michigan and Washington. CALDER Working Paper No. 247-1220
- Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, 1(1), 28-39.
- Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, 40(2), 190-222.
- Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2015). Evaluating specification tests in the context of value-added estimation. *Journal of Research on Educational Effectiveness*, 8(1), 35-59.
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*.
- Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180-195.
- National Council on Teacher Quality (2019). State of the States 2019. Policy Report. Washington, DC: National Council on Teacher Quality.
- Parsons, E., Koedel, C., & Tan, L. (2019). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics*, 44(2), 144-179.
- SAS Institute. (undated). SAS EVAAS: Statistical Models and Business Rules of TVAAS Analyses. Unpublished report.
- Vosters, K.N., Guarino, C.M., and Wooldridge, J.M. (2018). Understanding and evaluating the SAS® EVAAS® univariate response model (URM) for measuring teacher effectiveness. *Economics of Education Review* 66, 191-205.

Table 1. Descriptions of the five value-added specifications.

	Sparse		Student Controls		All Controls
	(1)	(2)	(3)	(4)	(5)
Structure	1-step	2-step	1-step	2-step	2-step
Student lagged test scores (math and ELA)	X	X	X	X	X
Individual student characteristics			X	X	X
School- and district-average student characteristics					X

Notes: All models also include fixed effects for student grade levels. The individual student characteristic controls are for race-ethnicity, gender, free/reduced-price lunch eligibility status, English language learner status, special education status, and mobility status. The school- and district-average characteristics are of these same variables, and lagged achievement, to control for the schooling environment.

Table 2. Summary statistics for students, schools, and districts in the analytic sample.

	Mean	Standard Deviation
Student Information		
Standardized math score	0.02	0.99
Standardized ELA score	0.02	0.99
Asian	0.02	0.14
Black	0.16	0.36
Hispanic	0.07	0.25
White	0.71	0.45
Multiple and Other Race-Ethnicity	0.04	0.20
Female	0.49	0.50
Eligible for free/reduced-price lunch	0.52	0.50
English language learner	0.05	0.22
Individualized education program	0.13	0.34
Mobile student	0.04	0.20
School Information		
Urban	0.18	0.38
Suburban	0.24	0.43
Rural/Town	0.59	0.49
Enrollment	357	217
District information		
Enrollment (all)	1603	3194
Avg. number of schools (all)	4.2	6.1
Enrollment (large district subsample)	6321	5333
Avg. number of schools (large district subsample)	12.4	11.0
N (student years, 2017-19)	972,877	
N (unique schools, 2017-19)	1,730	
N (unique districts, 2017-19)	557	

Notes: These summary statistics are based on the analytic sample of students in grades 4-8 in 2016-17, 2017-18, and 2018-19 who have lagged test scores and attend districts and schools with at least 10 test takers. Urbanicity information is taken from the 2018-19 Common Core of Data. The large-district subsample is selected to include the 100 districts in Missouri with the largest populations of test-takers included in the gap-year model. Other size-based selection criteria produce a similar sample; we chose this criterion in order to isolate districts in Missouri with the largest samples relevant for our primary analysis.

Table 3. Correlations between gap-year and full-data growth model output using different models and different data and estimation conditions.

		Math					ELA				
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 1	Model 2	Model 3	Model 4	Model 5
<u>District Models</u>											
	Baseline	0.91	0.91	0.90	0.90	0.90	0.88	0.88	0.88	0.89	0.90
	Same Cohorts	0.95	0.95	0.95	0.95	0.96	0.94	0.94	0.94	0.94	0.96
	Same Students	0.96	0.96	0.95	0.96	0.96	0.95	0.95	0.95	0.95	0.97
	Mobility Adjustment-1 (to baseline): omit movers	0.90	0.90	0.89	0.89	0.90	0.87	0.87	0.87	0.87	0.89
	Mobility Adjustment 2 (to baseline): 50-50 mover credit	0.90	0.90	0.90	0.90	0.91	0.88	0.88	0.88	0.88	0.90
<u>School Models</u>											
	Baseline	0.88	0.87	0.87	0.87	0.85	0.86	0.85	0.85	0.84	0.84
	Same Cohorts	0.91	0.92	0.91	0.92	0.90	0.89	0.89	0.89	0.89	0.89
	Same Students	0.91	0.92	0.91	0.92	0.90	0.89	0.89	0.89	0.89	0.90
	Mobility Adjustment-1 (to baseline): omit movers	0.88	0.86	0.87	0.85	0.84	0.86	0.83	0.86	0.83	0.83
	Mobility Adjustment 2 (to baseline): 50-50 mover credit	0.89	0.87	0.88	0.86	0.87	0.87	0.85	0.87	0.84	0.85

Notes: Each cell shows a correlation coefficient between growth measures using the gap-year and full-data scenarios.

Table 4. Observable predictors of changes to district growth rankings (in percentiles) due to the gap year in testing.

	Math					ELA				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 1	Model 2	Model 3	Model 4	Model 5
Period t-2 Test-Score Level (Same Subject)	-7.76* (0.75)	-7.52* (0.76)	-8.30* (0.81)	-8.06* (0.81)	2.36* (0.79)	-5.832* (0.817)	-5.406* (0.825)	-7.091* (0.838)	-6.737* (0.825)	2.021* (0.766)
Period t-2 Percent Asian	-0.18 (0.68)	-0.17 (0.68)	-0.13 (0.60)	-0.07 (0.59)	0.17 (0.81)	-0.525 (0.747)	-0.460 (0.781)	-0.316 (0.854)	-0.211 (0.852)	0.613 (0.770)
Period t-2 Percent Black	-1.37* (0.61)	-1.37* (0.61)	-2.23* (0.71)	-2.35* (0.70)	-0.57 (0.63)	-1.206* (0.605)	-1.194* (0.599)	-2.215* (0.630)	-2.148* (0.631)	0.056 (0.652)
Period t-2 Percent Hispanic	-0.68 (1.41)	-0.73 (1.46)	-1.04 (1.42)	-0.86 (1.48)	-0.58 (1.54)	-2.029 (1.377)	-2.082 (1.395)	-2.379 (1.392)	-2.318 (1.369)	-0.682 (1.365)
Period t-2 Percent Other Race	0.55 (0.61)	0.47 (0.61)	0.16 (0.63)	0.12 (0.63)	0.55 (0.70)	-0.496 (0.803)	-0.475 (0.807)	-0.467 (0.809)	-0.393 (0.801)	0.082 (0.755)
Period t-2 Percent female	-0.71 (0.61)	-0.67 (0.62)	-1.22* (0.65)	-1.21 (0.65)	-1.77* (0.76)	-0.014 (0.773)	0.045 (0.763)	0.184 (0.705)	0.188 (0.696)	-0.421 (0.721)
Period t-2 Percent FRL	-1.65* (0.80)	-1.75* (0.81)	-2.73* (0.88)	-2.99* (0.88)	1.07 (0.88)	-0.210 (0.854)	-0.278 (0.852)	-1.816* (0.856)	-2.307* (0.850)	0.176 (0.814)
Period t-2 Percent ESL	1.12 (1.40)	1.02 (1.45)	1.13 (1.43)	0.84 (1.49)	0.94 (1.67)	1.917 (1.469)	1.933 (1.498)	1.389 (1.473)	1.408 (1.454)	1.164 (1.389)
Period t-2 Percent IEP	-1.23* (0.49)	-1.15* (0.50)	-2.42* (0.65)	-2.33* (0.66)	0.46 (0.71)	-1.187* (0.517)	-1.066* (0.524)	-2.012* (0.543)	-1.904* (0.539)	0.660 (0.632)
Period t-2 Percent Mobile	-1.80* (0.65)	-1.75* (0.66)	-1.63* (0.71)	-1.65* (0.72)	0.80 (0.69)	-0.474 (0.722)	-0.379 (0.732)	-0.389 (0.718)	-0.367 (0.716)	1.926* (0.679)
Number of test takers	-0.14 (0.47)	-0.09 (0.48)	0.03 (0.48)	0.02 (0.47)	-0.14 (0.47)	-1.369* (0.468)	-1.428* (0.470)	-1.463* (0.542)	-1.420* (0.547)	-0.933 (0.502)
R-squared	0.247	0.229	0.242	0.233	0.054	0.160	0.140	0.176	0.158	0.039
N	540	540	540	540	540	540	540	540	540	540

Notes: The dependent variable in these regressions is each district's percentile ranking in the distribution of growth estimates using the gap-year data minus the percentile ranking using the full data. All variables are in standard deviations of the district distribution in period (t-2), which is 2017.

* Indicates statistical significance at the 5 percent level or higher.

Table 5. Observable predictors of changes to school growth rankings (in percentiles) due to the gap year in testing.

	Math					ELA				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 1	Model 2	Model 3	Model 4	Model 5
Period t-2 Test-Score Level (Same Subject)	-5.83* (0.82)	-5.41* (0.83)	-7.09* (0.84)	-6.74* (0.83)	2.02* (0.77)	-6.69* (0.67)	-5.91* (0.70)	-7.11* (0.69)	-6.44* (0.72)	1.46* (0.73)
Period t-2 Percent Asian	-0.53 (0.75)	-0.46 (0.78)	-0.32 (0.85)	-0.21 (0.85)	0.61 (0.77)	1.74* (0.57)	1.12 (0.58)	1.62* (0.57)	1.23* (0.59)	0.48 (0.62)
Period t-2 Percent Black	-1.21* (0.61)	-1.19* (0.60)	-2.22* (0.63)	-2.15* (0.63)	0.06 (0.65)	8.25* (2.79)	5.17 (3.12)	8.46* (2.79)	5.64 (3.11)	2.27 (3.15)
Period t-2 Percent Hispanic	-2.03 (1.38)	-2.08 (1.40)	-2.38 (1.39)	-2.32 (1.37)	-0.68 (1.37)	4.35* (1.37)	2.55 (1.45)	4.58* (1.39)	2.70 (1.46)	1.81 (1.45)
Period t-2 Percent Other Race	-0.50 (0.80)	-0.48 (0.81)	-0.47 (0.81)	-0.39 (0.80)	0.08 (0.76)	12.57* (3.02)	8.11* (3.37)	12.97* (3.01)	8.99* (3.35)	2.86 (3.39)
Period t-2 Percent female	-0.01 (0.77)	0.05 (0.76)	0.18 (0.71)	0.19 (0.70)	-0.42 (0.72)	0.08 (0.44)	-0.02 (0.47)	-0.25 (0.44)	-0.27 (0.47)	-0.53 (0.49)
Period t-2 Percent FRL	-0.21 (0.85)	-0.28 (0.85)	-1.82* (0.86)	-2.31* (0.85)	0.18 (0.81)	0.61 (0.61)	0.78 (0.66)	0.33 (0.63)	-0.59 (0.68)	1.29 (0.69)
Period t-2 Percent ESL	1.92 (1.47)	1.93 (1.50)	1.39 (1.47)	1.41 (1.45)	1.16 (1.39)	-1.65 (0.93)	-0.78 (0.88)	-2.33* (0.98)	-1.39 (0.93)	-1.11 (0.91)
Period t-2 Percent IEP	-1.19* (0.52)	-1.07* (0.52)	-2.01* (0.54)	-1.90* (0.54)	0.66 (0.63)	-1.82* (0.45)	-1.75* (0.45)	-2.55* (0.43)	-2.68* (0.46)	-0.20 (0.51)
Period t-2 Percent Mobile	-0.47 (0.72)	-0.38 (0.73)	-0.39 (0.72)	-0.37 (0.72)	1.93* (0.68)	-0.81 (0.47)	-0.40 (0.48)	-0.50 (0.54)	0.11 (0.55)	1.60* (0.60)
Number of test takers	-1.37* (0.47)	-1.43* (0.47)	-1.46* (0.54)	-1.42* (0.55)	-0.93 (0.50)	2.90* (0.32)	-0.17 (0.35)	3.03* (0.32)	-0.14 (0.35)	0.42 (0.37)
R-squared	0.160	0.140	0.176	0.158	0.039	0.141	0.100	0.159	0.097	0.013
N	540	540	540	540	540	1,527	1,527	1,527	1,527	1,527

Notes: The dependent variable in these regressions is each school's percentile ranking in the distribution of growth estimates using the gap-year data minus the percentile ranking using the full data. All variables are in standard deviations of the school distribution in period (t-2), which is 2017.

* Indicates statistical significance at the 5 percent level or higher.

Appendix A
Supplementary Tables

Appendix Table A1. Transition matrices documenting district and school ranking changes in the tails of the growth-ranking distributions. Baseline estimation conditions.

Panel A. Districts, Model 1

Math				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	6.9	3.1	0.0
	Middle 80 percent	3.1	73.3	3.5
	Top 10 percent	0.0	3.5	6.5

ELA				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	6.1	3.9	0.0
	Middle 80 percent	3.9	73.3	2.8
	Top 10 percent	0.0	2.8	7.2

Panel B. Districts, Model 2

Math				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	7.0	3.0	0.0
	Middle 80 percent	3.0	73.7	3.3
	Top 10 percent	0.0	3.3	6.7

ELA				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	5.9	4.1	0.0
	Middle 80 percent	4.1	72.8	3.1
	Top 10 percent	0.0	3.1	6.9

Panel C. Districts, Model 3

Math				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	6.1	3.9	0.0
	Middle 80 percent	3.9	72.8	3.3
	Top 10 percent	0.0	3.3	6.7

ELA				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	5.6	4.4	0.0
	Middle 80 percent	4.4	72.4	3.1
	Top 10 percent	0.0	3.1	6.9

Panel D. Districts, Model 4

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.5	3.5	0.0
	Middle 80 percent	3.5	73.1	3.3
	Top 10 percent	0.0	3.3	6.7

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	5.6	4.4	0.0
	Middle 80 percent	4.4	72.8	2.8
	Top 10 percent	0.0	2.8	7.2

Panel E. Districts, Model 5

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.7	3.3	0.0
	Middle 80 percent	3.3	73.9	2.8
	Top 10 percent	0.0	2.8	7.2

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.3	3.7	0.0
	Middle 80 percent	3.7	73.0	3.3
	Top 10 percent	0.0	3.3	6.7

Panel F. Schools, Model 1

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.1	3.9	0.0
	Middle 80 percent	3.9	73.2	2.9
	Top 10 percent	0.0	2.9	7.1

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.5	3.5	0.0
	Middle 80 percent	3.5	73.0	3.6
	Top 10 percent	0.0	3.6	6.4

Panel G. Schools, Model 2

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.7	3.3	0.0
	Middle 80 percent	3.3	73.5	3.2
	Top 10 percent	0.0	3.2	6.8

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.1	3.9	0.0
	Middle 80 percent	3.9	72.2	4.0
	Top 10 percent	0.0	4.0	6.0

Panel H. Schools, Model 3

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.2	3.8	0.0
	Middle 80 percent	3.8	73.5	2.8
	Top 10 percent	0.0	2.8	7.3

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.3	3.7	0.0
	Middle 80 percent	3.7	72.6	3.8
	Top 10 percent	0.0	3.8	6.2

Panel I. Schools, Model 4

Math				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	5.8	4.1	0.0
	Middle 80 percent	4.1	72.8	3.1
	Top 10 percent	0.0	3.1	6.9

ELA				
		Gap-year Data Growth Ranking		
Full Data Growth Ranking		Bottom 10 percent	Middle 80 percent	Top 10 percent
	Bottom 10 percent	6.2	3.7	0.0
	Middle 80 percent	3.7	72.6	3.7
	Top 10 percent	0.0	3.7	6.3

Panel J. Schools, Model 5

Math				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	6.0	3.9	0.0
	Middle 80 percent	3.9	72.8	3.3
	Top 10 percent	0.0	3.3	6.7

ELA				
		Gap-year Data Growth Ranking		
		Bottom 10 percent	Middle 80 percent	Top 10 percent
Full Data Growth Ranking				
	Bottom 10 percent	6.6	3.3	0.0
	Middle 80 percent	3.3	73.0	3.7
	Top 10 percent	0.0	3.7	6.4

Notes: Each cell in these tables indicates the percentage of Missouri districts for which the ranking profile matches the row and column headers. The sum of the cells in each matrix is 100 by construction (small discrepancies may arise due to rounding). Perfect alignment between the gap-year and full data estimates would result in diagonal cells of 10-80-10 with 0 values in all of the off-diagonal cells.

Appendix Table A2. Correlations between gap-year and full-data growth model output using different models and different data and estimation conditions. Large districts only.

	Math					ELA				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 1	Model 2	Model 3	Model 4	Model 5
<u>District Models (Large Districts Only)</u>										
Baseline	0.95	0.95	0.94	0.94	0.95	0.90	0.90	0.89	0.89	0.92
Same Cohorts	0.97	0.97	0.97	0.97	0.98	0.95	0.96	0.96	0.96	0.98
Same Students	0.97	0.97	0.97	0.97	0.99	0.96	0.96	0.96	0.96	0.98
Mobility Adjustment-1 (to baseline): omit movers	0.94	0.94	0.93	0.94	0.95	0.89	0.89	0.88	0.88	0.91
Mobility Adjustment 2 (to baseline): 50-50 mover credit	0.95	0.95	0.94	0.94	0.96	0.90	0.90	0.88	0.89	0.91

Notes: Each cell shows a correlation coefficient between growth measures using the gap-year and full-data scenarios.

Appendix Table A3. Correlations between gap-year and full-data growth model output using different models and different data and estimation conditions. 2-year gap scenario.

	Math					ELA				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 1	Model 2	Model 3	Model 4	Model 5
<u>District Models</u>										
Baseline	0.82	0.82	0.81	0.81	0.84	0.78	0.78	0.78	0.79	0.81
Same Cohorts	0.88	0.89	0.87	0.88	0.91	0.88	0.88	0.88	0.88	0.91
Same Students	0.89	0.90	0.88	0.89	0.93	0.89	0.89	0.89	0.90	0.93

Notes: Each cell shows a correlation coefficient between growth measures using the gap-year and full-data scenarios.

Appendix B

Explanation of the Gap-Year Effect on Growth Rankings

In this appendix we explain why the use of gap-year data in the less-specified VAMs (Models 1-4) results in lower growth rankings for districts and schools with higher academic achievement.

The explanation requires the presence of an omitted variable in the less-specified models that is positively correlated with students' lagged and contemporary test scores, which we treat as a static variable and refer to generally as the “schooling environment.” Our observational data do not allow us to test directly for the presence of such an omitted variable here, but evidence from Parsons, Koedel, and Tan (2019) indicates that it likely exists in these models.

In the equations that follow we show that, in the presence of such a variable, growth estimates from the sparsely specified models will be positively biased. In addition, the magnitude of the bias will be greater when growth over a two-year timespan is estimated based on consecutive single-year estimates—i.e., our full data condition—than when it is estimated over the same timespan but using just a single equation with a gap year in the test data—i.e., our gap-year data condition.

Importantly, this does not mean that the gap-year estimates are preferred to estimates based on the full data from underspecified VAMs because the gap-year estimates have other limitations (most notably, worse coverage and smaller samples). However, under the reasonable assumption that the biasing variable—i.e. the “schooling environment”—is positively correlated with the achievement level in schools and districts, it explains the systematic, negative association between the baseline achievement level and the change in growth rankings caused by the gap year in the underspecified VAMs.

First, consider the following single-year growth model, which is a simplified version of the first-stage of Model 2 in the text:¹⁹

$$Y_{ist} = \beta_0 + Y_{ist-1}\beta_1 + X_{is}\alpha + e_{ist} \tag{B1}$$

In equation (B1), Y_{ist} is the standardized achievement of student i who attends school s in year t , X_{is} is a variable that captures the unobserved “schooling environment” at the school attended by

¹⁹ In writing out these equations, we assume a simple model that controls just for lagged achievement in the same subject (such that Y_{ist-1} is a scalar). This simplification has no bearing on the substance of this appendix. The insights also apply to one-step VAMs, although we use the two-step structure to illustrate.

student i , which is assumed to be time invariant over the modeling period, and e_{ist} is the error term. Because X_{is} is unobserved, when we estimate this model with available data, we estimate:

$$Y_{ist} = \beta_0 + Y_{ist-1}(\beta_1 + \alpha\delta_1) + u_{ist} \quad (B2)$$

Like terms in equation (B2) are defined as in equation (B1), and note that δ_1 is from the regression: $X_{is} = \delta_0 + Y_{ist-1}\delta_1 + \mu_{is}$. Thus, the term $\alpha\delta_1$ indicates the degree of bias in the estimated coefficient on Y_{ist-1} resulting from the omitted variable.

Given the above, the residuals from the true model (equation B1) can be written as:

$$R(true)_{ist} = Y_{ist} - \{\beta_0 + Y_{ist-1}\beta_1 + X_{is}\alpha\} \quad (B3)$$

and the residuals from estimated model (equation B2) can be written as:

$$R(est)_{ist} = Y_{ist} - \{\beta_0 + Y_{ist-1}(\beta_1 + \alpha\delta_1)\} \quad (B4)$$

Taking the difference of these residuals yields the following equation, which indicates the degree to which student residuals produced by the estimated model will differ from those produced by the true model.

$$R(est)_{ist} - R(true)_{ist} = \alpha(X_{is} - Y_{ist-1}\delta_1) \quad (B5)$$

Because the school and district growth estimates are aggregations of the student residuals in the two-step VAMs, equation (B5) forms the basis of the differences between the gap-year and full-data growth estimates. The process of aggregating the values from equation (B5) to the district or school levels produces the differences in the growth estimates reported in the main text.

Moving to the gap-year model, note that substituting a version of equation (B1) where Y_{ist-1} is the dependent variable into itself (i.e., equation B1) produces the following gap-year model analog:

$$Y_{ist} = \gamma_0 + Y_{ist-2}\gamma_1 + X_{is}\tilde{\alpha} + \eta_{ist} \quad (B1')$$

where $\gamma_0 = \beta_0(1+\beta_1)$, $\gamma_1 = (\beta_1)^2$, $\tilde{\alpha} = \alpha(1 + \beta_1)$, and $\eta_{ist} = (1 + \beta_1)e_{ist}$. As a result, the corresponding gap-year analog of equation (B5) can be written as:

$$R(est)_{ist} - R(true)_{ist} = \tilde{\alpha}(X_{is} - Y_{ist-2}\delta_1) \quad (B5')$$

Given the above, we can specify the following function $Z(X)$, which indicates how the level of bias in the student residuals differs when comparing the full-data model to the gap-year model.

$$Z(X) = \{\alpha(X_{is} - Y_{ist-1}\delta_1) + \alpha(X_{is} - Y_{ist-2}\delta_1)\} - \{\tilde{\alpha}(X_{is} - Y_{ist-2}\delta_1)\} \quad (B6)$$

The first two terms in equation (B6) result from the fact that each student has two residuals in the full-data model – one from time $t-1$ and one from time t . Meanwhile, the third term is from the gap-year model, in which each student has only one residual.

Taking the derivative of (B6) with respect to X yields

$$Z'(X) = \{\alpha - Y'_{ist-1}(X)\alpha\delta_1\} + \{\alpha - Y'_{ist-2}(X)\alpha\delta_1\} - \{\tilde{\alpha} - Y'_{ist-2}(X)\tilde{\alpha}\delta_1\} \quad (\text{B7})$$

Next, we make the simplifying assumption that Y_{ist-2} is the first year in which exams are taken and replace the lagged exam score with an ability endowment, Y_0 , standardized to be on the same scale as the exam and independent of X , i.e. $Y_{ist-2} = \beta_0 + Y_0\beta_1 + X_{is}\alpha + e_{ist-2}$.²⁰ Then,

$$Y'_{t-2}(X) = \alpha$$

$$Y'_{t-1}(X) = \alpha(1+\beta_1)$$

Substituting these expressions into (B7) and simplifying, again noting that $\tilde{\alpha} = \alpha(1 + \beta_1)$, yields:

$$Z'(X) = \alpha(1 - (\alpha\delta_1 + \beta_1)) \quad (\text{B8})$$

If $\alpha > 0$ and $(\alpha\delta_1 + \beta_1) < 1$, then the derivative in (B8) is positive, which means that the gap in the student residuals between the full-data and gap-year models is increasing in X . The first of these conditions is intuitive and supported by evidence in Parsons, Koedel, and Tan (2019), while the second condition is widely shown in empirical research and confirmed in our data (note that the second term is just the coefficient on the lagged test score estimated by the feasible model shown in equation B2).

In summary, these equations show that systematic, positive bias in district and school growth estimates from underspecified single-year VAMs is fully compounded across years when single-year estimates are combined to estimate growth over two years. Bias in the same direction exists in the gap-year model, but its impact over the two years is attenuated by the modeling structure. Again, this does not mean that gap-year estimates are preferable to full-data estimates from underspecified VAMs because the gap-year estimates have other limitations, most notable in terms of coverage and sample sizes. It is also worth noting that the magnitudes of bias involved in these equations are likely modest based on previous research. Still, it does explain

²⁰ This simplification is to make the mathematics tractable. In practical terms, we are assuming that the omitted variable X is the product of a true effect of exposure to the schooling environment and not due to student sorting at entry. The substance of what we show here does not depend on this assumption—and the bias will only be made more pronounced by additional sorting bias in the same direction—but it greatly simplifies the expressions that follow.

the directional results in Tables 4 and 5, which show the gap-year models systematically lower the growth rankings of high-achieving schools and districts.

This framework also helps to reinforce the idea that growth estimates from the fully-specified VAM (Model 5) are less biased than from the other specifications, at least to the extent that the bias is correlated with observable characteristics. This inference comes from (a) the equations above showing that estimates from the gap-year analog to *any* full-data model will be less biased, and (b) our finding that growth-ranking changes caused by the gap year in Model 5 are not meaningfully related to measurable district or school characteristics (per Tables 4 and 5 in the paper). A critique of the fully-specified model is that it may “overcorrect” for student and school circumstances (Ehlert, Koedel, Parsons, and Podgursky, 2016), and this could also generate our results for that model in the main text. However, findings from Parsons, Koedel, and Tan (2019) suggest that the bias in the other direction is likely more important.